Building Text-to-Speech Systems for Resource Poor Languages

Nur-Hana Samsudin, Mark Lee

School of Computer Science,
University of Birmingham, Birmingham, UK
{*n.h.samsudin, m.g.lee*}*@cs.bham.ac.uk*

Abstract

This paper describes research on building text-to-speech synthesis systems (TTS) for resource poor languages using available resources from other languages and describes our general approach to building cross-linguistic polyglot TTS. Our approach involves three main steps: language clustering, grapheme to phoneme mapping and prosody modelling. We have tested the mapping of phonemes from German to English and from Indonesian to Spanish. We have also constructed three prosody representations for different language characteristics. For evaluation we have developed an English TTS based on German data, and a Spanish TTS based on Indonesian data and compared their performance against pre-existing monolingual TTSs. Since our motivation is to develop speech synthesis for resource poor languages, we have also developed three TTS for Iban, an Austronesian language with practically no available language resources, using Malay, Indonesian and Spanish resources.

Keywords: TTS, polyglot speech synthesis, resource adaptation, multilinguality

1. Introduction

This paper focuses on the creation of polyglot text-to-speech (TTS) systems. We distinguish polyglot and multilingual TTS as follows: a multilingual TTS has different algorithms, rules and speech data for different languages, while in a polyglot TTS, there is a primary language which is the focal language of the synthesiser (Traber et al., 1999). The main advantage of polyglot speech synthesis is that a system using the polyglot framework will be able to synthesise multiple languages using the same recorded or trained voices.

In the polyglot architecture, all the different language resources are combined to train the voices so that the system can retain the wave signal for reproduction during synthesis (Tokuda et al., 2002). This approach requires the target language's data to be accessible during training.

However, when such data is not available, or when a language is resource poor, it is difficult to construct a polyglot TTS. Therefore, in this paper, we propose a novel method of reusing available resources using our proposed linguistic characterisation. This paper covers the adaptation of phonemes and the reusing of another language's voices to create a TTS for a resource poor language. In addition, the approach presented in this paper is suitable for rapid TTS prototyping for better resourced languages.

Throughout this paper, we refer to the language where the data originates from as the native language and the language generated using the adapted data as the target language. Throughout this paper, the proof-of-concept we refer to uses the multilingual voice database MBROLA (MBROLA-Group, 2005). The data used in our approach is the voice recordings of the native language and its phoneme set. Finally, the prosody of the polyglot TTS is constructed using our proposed prosody representation.

This paper is organised into six sections. Section 2 presents our reuse methodology for different language resources and the studies conducted on multiple languages. Section 3 describes the approach taken to adapt from one language to another using the proposed polyglot framework. Section 4 summarises the creation of an Iban TTS using resources from several languages. Section 5 lists the results and evaluation summary conducted on the English, Spanish, Malay and Iban polyglot TTSs. Section 6 will conclude the overall paper and describe future work.

2. Brief Description on Reusing Different Language Resources for Polyglot TTS

To test the adaptability of resource poor languages within the framework, three sets of pilot studies were conducted. The first set involved the creation of an English Polyglot TTS using German data, the second set involved creating a Spanish polyglot TTS using Indonesian data and finally, the third set involved the creation of a Malay Polyglot TTS using Afrikaans and English data. These will be further described below.

2.1. English Polyglot TTS

An English polyglot TTS was constructed based on the German MBROLA phoneme set and voices. There are 55 phonemes identified in the German phoneme list, and 44 in the English data (MBROLA-Group, 2005). The mapping is done based on the SAMPA representation and then checked manually to confirm the mapping's validity. Then the grapheme-to-phoneme (G2P) mapping and prosody is implemented based on our proposed mapping and prosody representation.

2.1.1. German Phonemes to English Phonemes

Table 1 shows the list of English phonemes and the corresponding phonemes in German where the phoneme exists. When a phoneme of the target language does not exist in the native language, the substituted phoneme is selected from the phoneme substitution matrix (This is described further in Section 3.2.).

Based on Table 1, thirteen English phonemes are missing from the German data. To compensate for these missing

Table 1: Adapting English phonemes from German phonemes using (MBROLA-Group, 2005)

English IPA (SAMPA)	Matched German IPA (SAMPA)	English IPA (SAMPA)	Matched German IPA (SAMPA)
p (p)	p (p)	w (w)	w (w)
b (b)	b (b)	l (l)	l (l)
t (t)	t (t)	ɪ (I)	ɪ (I)
d (d)	d (d)	ʊ (U)	ʊ (U)
k (k)	k (k)	e (e)	ɛ (E)
g (g)	g (g)	ə (@)	e (e)
m (m)	m (m)	æ ({)	a+ɛ (a+E)
n (n)	n (n)	ʌ (V)	a (a)
ŋ (N)	ŋ (N)	ɒ (Q)	ɔ (O)
r (r)	r (r)	iː (i:)	iː (i:)
f (f)	f (f)	uː (u:)	uː (u:)
v (v)	v (v)	ɔː (O:)	oː (o:)
θ (T)	θ (T)	ɜː (3:)	œ (9:)
ð (D)	ð (D)	ɑː (A:)	aː (a:)
s (s)	ð (D)	eɪ (eI)	ɛɪ (EI)
z (z)	z (z)	aɪ (aI)	aɪ (aI)
ʃ (S)	ʃ (S)	ɔɪ (OI)	ɔʏ (OY)
ʒ (Z)	ʒ (Z)	əʊ (@U)	əʊ (@U)
h (h)	h (h)	aʊ (aU)	aʊ (aU)
t͡ʃ (tS)	t͡ʃ (tS)	ɪə (I@)	ɪ+ə (I+@)
d͡ʒ (dZ)	ç (C)	eə (e@)	ɛ+ə (E+@)
j (j)	j (j)	ʊə (U@)	ʊ+ə (U+@)

phonemes, a phoneme substitution matrix has been constructed consisting of a set of proposed phoneme substitutions for frequently used phonemes in most languages. These substitution phonemes are proposed based on the closest similarity of manner and place of articulation for pulmonic consonants and position of the tongue and roundedness for vowels. The English substituted phonemes based on German data are listed as follows in IPA and then SAMPA in parenthesis: /d͡ʒ/→/ç/, /ɑː/→/aː/, /ɔː/→/oː/, /ɜː/→/œ/, /e/→/ɛ/, /æ/→/a+ɛ/, /ʌ/→/a/, /ɒ/→ /ɔ/, /eɪ/→/ɛɪ/, /ɔɪ/→/ɔʏ/, /ɪə/→/ɪ+ə/, /eə/→ /ɛ+ə/ and /ʊə/→/ʊ+ə/. The phoneme /ɒ/ does exist in the German data, however it is categorised as a controlled phoneme in which the phoneme can be used only in a specific sequence of phonemes. The phoneme /ɔ/ is used as the substituted phoneme. English phoneme: /d͡ʒ/ does not exist in German data. The next three best substituted phonemes for it are: /d͡z/, /t͡ʃ/ and /t͡s/. However, /t͡ʃ/ cannot be selected as the substitution because English also uses this phoneme. The closest sound-based on the IPA chart would be: /d͡zˌ/ and /t͡sˌ/ but it is not available. Due to that, the available close sounding phoneme is selected: /ç/. It is worth noting that the native language's data which had been selected catered for a few phoneme only available in loan words, namely: /θ/ and /ð/. The grapheme-to-phoneme (G2P) for English is based on the Longman pronunciation dictionary. In addition, irregular orthographic mapping is used for proper nouns and words that are not converted by the pronunciation dictionary.

2.1.2. English Prosody

English is a stressed language where the stressed location is not fixed. For the English polyglot study, we used the Longman pronunciation dictionary to mark the stresses and this information is passed on for prosody processing. The prosody is then assigned following the stressed language prosody.

2.2. Spanish Polyglot TTS

The Spanish polyglot TTS was constructed based on the Indonesian MBROLA phoneme set and voices. There are 34 phonemes listed in the Spanish phoneme list, and 29 phonemes in the Indonesian data (MBROLA-Group, 2005). The mapping is done based on the SAMPA representation and is checked and updated manually then the grapheme-to-phoneme mapping and prosody is implemented based on our proposed prosody representation.

2.2.1. Indonesian Phoneme to Spanish Phoneme

Although there are 34 phonemes used in Spanish, these have been modified in the MBROLA Spanish data set, with the effect that the SAMPA symbols used do not tally with standard SAMPA. This was perhaps originally done to make it easier for the original voice developer to assign the prosody in a Spanish monolingual TTS system. To ensure the uniformity of the polyglot data, we have modified the Spanish phoneme list to follow the standard SAMPA notation.

Table 2: Adapting Spanish phonemes from Indonesian phonemes(MBROLA-Group, 2005)

Spanish IPA (SAMPA)	Matched Indonesian IPA (SAMPA)	Spanish IPA (SAMPA)	Matched Indonesian IPA (SAMPA)
p (p)	p (p)	h (h)	h (h)
b (b)	b (b)	t͡ʃ (tS)	t͡ʃ (tS)
t (t)	t (t)	d͡ʒ (L)	d͡ʒ (dZ)
d (d)	d (d)	j (j)	j (j)
k (k)	k (k)	w (w)	w (w)
gˈ (g)	gˈ (g)	r (r)	r (r)
g (g)	g (g)	ɾ (4)	rː(r:)
m (m)	m (m)	l (l)	l (l)
n (n)	n (n)	a (a)	ʌ (V)
ɲ (J)	ɲ (J)	e (e)	e (e)
β (B)	b (b)	o (o)	ɒ (Q)
f (f)	f (f)	i (i)	ɪ (I)
ð (D)	d+h (d+h)	u (u)	ʊ (U)
s (s)	s (s)		

The Spanish polyglot synthesiser is developed using Indonesian data as an example of adaptation between different language families. The Spanish-Indonesian phonemes are listed in Table 2. From this list, we find that the following phonemes in Spanish data are not available in the Indonesian data and therefore we listed the substitution based on the phoneme substitution matrix: /ɾ/ → /rː/, /a/ → /ʌ/, /ð/ → /d+h/, /i/ → /ɪ/, /o/ → /ɒ/ and /u/ → /ʊ/. The G2P

for Spanish uses our modified phonemic orthographic mapping. This set of rules is selected because Spanish is a phonemic orthographic language.

2.2.2. Spanish Prosody

Spanish is a fixed stress language. Therefore we use linguistic rules to identify the stress position. Spanish also follows our proposed prosody template for fixed stress location.

2.3. Malay Polyglot TTS

Two Malay polyglot TTSs were constructed using two different language resources: English and Afrikaans. There are 34 phonemes listed in the Malay phoneme list; 37 phonemes in Afrikaans and 44 phonemes in English (MBROLA-Group, 2005). The phoneme mappings are matched based on the SAMPA representation and then checked and updated manually. The G2P mapping and prosody are then assigned based on our proposed phonemic orthographic mapping and prosody representation respectively.

2.3.1. English and Afrikaans Phonemes to Malay Phonemes

The set of Afrikaan's phonemes are lacks several Malay phonemes and therefore the substitute phonemes for resource from Afrikaans are: /əʊ/→/ə+u/, /aɪ/→/a+i/, /aʊ/→/a+w/, /dʒ/→/d+ʒ/, /eɪ/→/e+i/, /ɪ/→/i/, /ɲ/→/n+j/, /ɔɪ/→ /oi/, /ɒ/→/ɔ/, /tʃ/→/t+ʃ/, /ʌ/→/a/, /ʊ/→/u/. Although the /ʌ/ is actually closer to /ɑ/ in the IPA chart, due to the limitations of /ɑ/ usage in Afrikaans, we use /a/ in the polyglot TTS.

Comparing between the English phoneme list in Table 3 in column two, only the following phonemes are not available in the English phoneme list and therefore substituted: /ɲ/→/n+j/ and /x/→/k+h/. Based on the number of phoneme substitutions required, the Malay phonemes are a better match with English than with Afrikaans although based on the accent inside each recorded file itself, Afrikaans is more closely related to Malay.

Table 3: Adapting Malay phonemes from English and Afrikaans phonemes (MBROLA-Group, 2005)

Malay IPA (SAMPA)	Matched English IPA(SAMPA)	Matched Afrikaans IPA (SAMPA)
p (p)	p (p)	p (p)
b (b)	b (b)	b (b)
t (t)	t (t)	t (t)
d (d)	d (d)	d (d)
k (k)	k (k)	k (k)
g (g)	g (g)	g (g)
m (m)	m (m)	m (m)
n (n)	n (n)	n (n)
ŋ (N)	ŋ (N)	ŋ (N)
ɲ(J)	n+j (n+j)	n+j (n+j)
r (r)	r (r)	r (r)
f (f)	f (f)	f (f)
s (s)	s (s)	s (s)
z (z)	z (z)	z (z)
ʃ (S)	ʃ (S)	ʃ (S)
x (x)	k+h (k+h)	x (x)
h (h)	h (h)	h (h)
dʒ (dZ)	dʒ (dZ)	d+ʒ (d+Z)
tʃ (tS)	tʃ (tS)	t+ʃ (t+S)
j (j)	j (j)	j (j)
w (w)	w (w)	w (w)
l (l)	l (l)	l (l)
ʌ (V)	ʌ (V)	ɑ (A)
ə (@)	ə (@)	ə (@)
e (e)	e (e)	e (e)
ɪ (I)	ɪ (I)	i (i)
ɒ (Q)	ɒ (Q)	ɔ (O)
ʊ (U)	ʊ (U)	u (u)
aɪ (aI)	aɪ (aI)	a+i (a+i)
aʊ (aU)	aʊ (aU)	a+w (a+w)
eɪ (eI)	eɪ (eI)	e+i (e+i)
əʊ (@U)	əʊ (@U)	ə+u (@+u)
ɔɪ (OI)	ɔɪ (OI)	o+i (o+i)

2.3.2. Malay Prosody

Malay is not a stressed or tonal language, so there is no objectively correct way to pronounce a word (Don et al., 2008). However certain tones may sound unnatural.

3. Adaptation Process from Native Language to Target Language

In this section we will describe the process flow for adapting from a Native to a Target Language. In the next Section we describe phoneme matching and then present a G2P mapping template for phonetic mapping. Finally we will briefly describe the prosody template used.

3.1. Phoneme Matching

In the adaptation process, phonemes of the target language may not exist in the native language, therefore phoneme substitution is required. When there is no match among the phonemes in the phoneme substitution matrix, substitution is selected based on the closest sounding phoneme available. The phoneme substitution matrix is classified into vowels, diphthongs and consonants. Due to space restrictions, we provide just a sample of the phoneme substitution matrix.

3.2. Grapheme-to-Phoneme Mapping

G2P conversion is the first process performed on normalised text. We have created two sets of global phoneme mappings. One is for regular orthographic mappings and another one is for irregular orthographic mapping. Languages like Spanish, Italian, German, Indonesian and Malay are categorised as phonemic orthographic languages. English RP and French are both categorised as irregular orthographic languages. For each polyglot TTS, only one mapping is applicable.

3.2.1. Phonemic Orthographic Mapping

Table 5 shows the general phonemic orthographic mapping. For each of the following language: Spanish, Malay and Iban, this mapping is further refined to match with the target

Table 4: Snippet from Phoneme Substitution Matrix

The IPA	1st Sub	2nd Sub	3rd Sub
i	ɪ	i̞	
ɪ	i	i̞	ʏ
u	ɯ	ʊ	
ʊ	ɯ	u	
e	ɛ		
o	ɔ	ɒ	
ə	ɜ	ɘ	ɞ
ʌ	ɑ	a	
a	ɐ	æ	ɑ
ɑ	ʌ		
ɾ	r	ɽ	
j̠	n+j		
x	ʍ	k+h	
t͡s	c	t͡ʃ	t+s
d͡ʒ	d͡z	d͡ʐ̩	d͡ʑ

language.

Table 5: Basic Phonemic Orthographic Mapping

Grapheme	Phoneme	Grapheme	Phoneme
Vowel			
a	a	o	o
e	ə	u	u
i	i		
Consonants			
b	b	p	p
c	t͡ʃ	q	q
d	d	r	r
f	f	sh	ʃ
g	g	sy	ʃ
h	h	s	s
j	d͡ʒ	t	t
k	k	v	v
l	l	w	w
m	m	x	ks
ng	ŋ	y	j
ny	ɲ	z	z
n	n		

3.2.2. Irregular Orthographic Mapping

Irregular orthographic mapping uses language dependent rules where phoneme sequences are based on specific sequences of graphemes. The irregular orthographic mapping constructed is based on the English G2P mapping. Table 6 provide the most common English grapheme to phoneme conversion. For the English polyglot TTS, the G2P conversion uses the Longman pronunciation dictionary (PhoTransEdit-Group, 2011).

3.3. Prosody Representation

Three things need to be identified before prosody assignment: language clustering, global phoneme mapping and intonation type identification. In language clustering, the clusters used for language categorisation are introduced. These include determining whether the language is orthographical, stressed or unstressed language and tonal or non-tonal. After checking the orthographic aspect, the weight of the most frequently used phonemes of the target language is identified and compared to the possible 'adopted' language.

Table 6: Irregular Orthographic Mapping based on the most common G2P in English (summarise in Marks (2011))

Grapheme	Phoneme	Grapheme	Phoneme
Vowel			
C+a+C	æ	e+w	juː
a+C+e	eɪ	i+C+e	aɪ
a+i	eɪ	i+e+_	aɪ
a+i+r	e+ə	i+e	iː
ar	ɑː	i+r	ɜː
au	ɔː	i	ɪ
aw	ɔː	o+C+e	əʊ
ay	eɪ	o+a	əʊ
a	ə	o+i	ɔi
_+e+a+r	ɜː	o+o+r	ɔː
e+C	e	o+o	ʊ
e+C+e	iː	o+r	ɔː
C+e+C+_	ə	o+u+n	aʊ
C+ea+_	iː	o+u+l	ʊ
e+a+r+_	eə	o+u+C	uː
e+r+e+_	eə	o+u+gh	ɔː
e+n+g	ɪ	o+w	aʊ
C+ea+C	e	o+y	ɔɪ
e+a+r	ɪə	o	ɒ
e+e	iː	ur	ɜː
e+e+r	ɪə	eu	uː
ei	eɪ	C+u+C+_	ʌ
ey	eɪ	u	ʊ
Consonants			
b	b	p	p
c+e/i/y	s	qu	kw
c+h	t͡ʃ	q	k
c+k	k	r	r
c+C+e	s	s+y	z+i
c	k	sh	ʃ
d	d	V+s+u	ʒ
f	f	s	s
g+V+C+_	g	tch	t͡ʃ
g+u+e+C	g+e	V+t+V	ʃ
g+u	g+w	c/k+t+u	t͡ʃ
g	g	th	θ
h	h	t	t
j	d͡ʒ	v	v
k	k	wh	w
l	l	w	w
m+u	mjuː	C+C+y+_	aɪ
m	m	y+C+e	aɪ
n+c/k	ŋ	_+y	j
ng	ŋ	y+_	i
n	n	z	z
ph	f		

In prosody assignment, the prosody value is calculated and assigned to each sound unit. We calculate this based on the International Symbol of Intonation (INTSINT) proposed by Hirst (2005) & (2007). The calculation of each symbol is provided by the following equations:

$$\mathbf{T}op = key * \sqrt{2^{range}} \quad (1)$$

$$\mathbf{M}edium = key \quad (2)$$

$$\mathbf{B}ottom = key/\sqrt{2^{range}} \quad (3)$$

$$\mathbf{H}igher = \sqrt{P_{i-1} * T} \quad (4)$$

$$\mathbf{U}pperstep = \sqrt{P_{i-1} * \sqrt{P_{i-1} * T}} \quad (5)$$

$$\mathbf{S}ame = P_{i-1} \quad (6)$$

$$\mathbf{D}ownstep = \sqrt{P_{i-1} * \sqrt{P_{i-1} * B}} \quad (7)$$

$$\mathbf{L}ower = \sqrt{P_{i-1} * B} \quad (8)$$

The **T, M, B** values are absolute values where the value of the current pitch is not influenced by any preceding pitch. **H, U, S, D** and **L** tones are defined by the preceding pitch value. P_{i-1} refers to the previous pitch. The mappings depicted in Figure 1 is used to map the INTSINT labelling correspondences. To describe the pitch contours of languages, the movement of pitch is described based on the pitch points shown in Figure 1.

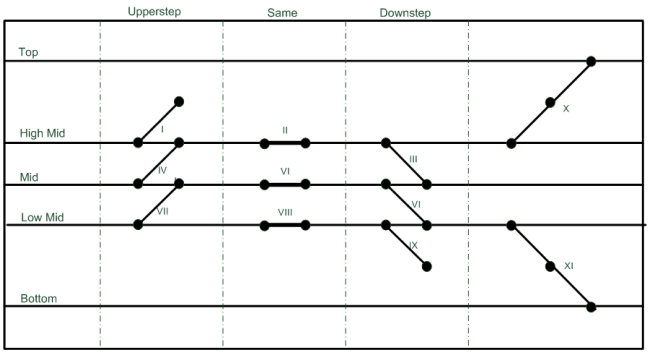

Figure 1: The pitch points movement for stressed language and open intonation language.

Prosody manipulation is different for stressed languages (English and Spanish) and open intonation languages (Malay). For a stressed language, the following pitch contours labelled and descriptions are as follows: Syllable with the lines I, IV and VII holds the previous pitch's value or INTSINT value; **M**edium; and at the end syllable, the pitch holds the **U**pperstep value. For the lines II, V and VIII, no changes of initial and final pitch of the syllable occur. For the lines III, VI and IX, the final values will hold the **D**ownstep value. While for the lines X and XI, the calculation of **U**pperstep and **D**ownstep will be calculated twice respectively. For every stressed syllable, the syllable duration is extended. Then the first phoneme of the next syllable will be assigned to **D**ownstep so that the stressed syllable is prominent.

For an open intonation language, the pitch contour can be represented similar to Figure 1. However, the rising and falling of a syllable is subjective. Thus, for open intonation language, the tone is changed for every other syllable in a word. Therefore the pitch will be changed at every first and third syllable unless for a word immediately follows by punctuation; comma, full stop and colon; the pitch will be assigned to the final syllable of the word.

4. Creating Iban Polyglot TTS

Iban is a language within the Austronesia family and more specifically is related as follows: Malayo-Polynesian → Malayo-Sumbawan → Malayic → Iban (Lewis, 2011). It is spoken in Brunei, the Sarawak state of Malaysia and West Kalimantan Province of Indonesia. Crystal (2010, p. 471) estimates that the language is spoken by 561,000 people.

The Iban phoneme set was developed by researchers from Sarawak Language Technology (SaLT laboratory of Universiti Malaysia Sarawak (UniMAS). The G2P and the phonological rules of Iban have been developed together with this collaboration. Iban grapheme-to-phoneme is modified from the phonemic orthographic mapping. The phoneme sequence is compared to the native speaker's speech for validation. Iban is a language closely related to Malay (Maynes and Gara, 2011). It can be categorised into two: formal Iban and informal Iban (Maynes and Gara, 2011). Both Iban are used in Jako Iban which is a conversational language (Saee, 2011). Iban is an open intonation language. Thus the ideal prosody template for an Iban polyglot TTS is the open intonation template.

The Iban polyglot TTS were implemented using three different language resources. First is Malay using Malay sound files. Second is the Indonesian resource using the open intonation template introduced earlier. The third is the Iban polyglot TTS using Spanish resources and the stressed language prosody template. However, for detailed evaluation, only the Malay and Indonesian resources were used.

From Table 7, one critical vowel which is used extensively in Iban language is /ə/ but was changed to /e/ for Spanish resources even when the Iban language itself requires the /e/ sound. Other vowels required in Iban are mapped based on the closest AEIOU sound available in the other languages: Malay and Indonesian.

5. Survey and Results

For the English TTS, twenty-two native English speakers participated in the survey and a total of 110 evaluations were obtained for the set of short sentences and long sentences. In total, the respondents needed to listen to 10 sets of synthesised speech where in each set, there are two wave files. For the Spanish TTS evaluation, ten native Spanish speakers participated in the survey and 46 evaluations were obtained for the set of short sentences and 50 evaluations for long sentences. Thirty respondents participated in the Malay survey. The evaluations of the polyglot speech are all compared to the monolingual TTSs.

In the English and Spanish evaluation, a polyglot TTS system was compared to a monolingual TTS for each language. For Malay, three synthesisers were developed using the proposed framework. The Malay TTSs were constructed using data from Malay itself, Afrikaans and English. The Iban polyglot TTSs were constructed using data

Table 7: Adapting Iban phonemes from Malay, Indonesian and Spanish phonemes (MBROLA-Group, 2005)

Iban IPA (SAMPA)	Matched Malay IPA (SAMPA)	Matched Indonesian IPA (SAMPA)	Matched Spanish IPA (SAMPA)
p (p)	p (p)	p (p)	p (p)
b (b)	b (b)	b (b)	b (b)
t (t)	t (t)	t (t)	t (t)
d (d)	d (d)	d (d)	d (d)
k (k)	k (k)	k (k)	k (k)
g (g)	g (g)	g (g)	g (g)
ʔ (?)	(silent)	(silent)	(silent)
m (m)	m (m)	m (m)	m (m)
n (n)	n (n)	n (n)	n (n)
ŋ (N)	ŋ (N)	ŋ (N)	n+g (n+g)
ɲ(J)	ɲ(J)	ɲ(J)	ɲ(J)
r (r)	r (r)	r (r)	r (r)
s (s)	s (s)	s (s)	s (s)
z (z)	z (z)	z (z)	z (z)
h (h)	h (h)	h (h)	h (h)
dʒ (dZ)	dʒ (dZ)	dʒ (dZ)	dʒ (dZ)
tʃ (tS)	tʃ (tS)	tʃ (tS)	tʃ (tS)
j (j)	j (j)	j (j)	j (j)
w (w)	w (w)	w (w)	w (w)
l (l)	l (l)	l (l)	l (l)
ɑ (A)	ʌ (V)	ʌ (V)	a (a)
ə (@)	ə (@)	ə (@)	e (e)
e (e)	e (e)	e (e)	e (e)
i (i)	ɪ (I)	ɪ (I)	i (i)
ɔ (O)	ɒ (Q)	ɒ (Q)	o (o)
ɤ (7)	ɒ (Q)	ɒ (Q)	o (o)
ʊ (U)	ʊ (U)	ʊ (U)	u (u)
ɑi (Ai)	aɪ (aI)	aɪ (aI)	a+j (a+j)
aʊ (aU)	aʊ (aU)	aʊ (aU)	a+u (a+u)

from Malay, Indonesian and Spanish. The result of the English, Spanish and Malay evaluations are presented in the following table.

Table 8: Overall Comparative Quality and Naturalness Rating between polyglot and molingual TTS

	English			Spanish			Malay		
	M	P	Un.	M	P	Un.	M	P1	P2
Clarity	75	19	6	98	0	2	76	17	5
Natural	48	47	5	98	0	2	26	41	26

At the initial stage of evaluation, the respondents were asked to rate the overall quality and naturalness of the polyglot speech. For the English and Spanish sub column of Table 8 and 9, the **M**, **P** and **Un.** stands for monolingual, polyglot TTS and uncertain respectively. The subcolumn of Malay: **M**, **P1** and **P2** mean monolingual, the first polyglot TTS (using English data) and the second polyglot TTS (using Afrikaans data) respectively. The left-most column in Table 9 and 12 refers to the perceived effort that is required to understand the synthesised speech rated on a three point scale from least to considerable and "none" where no meaning was understood and "uncertain" where the subject was uncertain. Table 8 to Table 13 shows the respondents feedback in percentage values.

Table 9: Overall effort comparison rating between polyglot and monolingual TTSs

	English		Spanish		Malay		
	M	P	M	P	M	P1	P2
Least	32	9	49	25	60	26	15
Moderate	47	37	47	44	29	42	34
Considerable	16	43	4	29	9	27	42
None	4	10	0	2	2	5	9
Uncertain	1	1	0	0	0	0	0

5.1. Summary of English, Spanish and Malay Results

It is expected that the respondent will prefer the monolingual TTS speech over a polyglot synthesiser. Table 10 to 13 shows the results of evaluations where the respondents are requested to listen, type and rate the sound that they hear from a series of meaningless sentence.

Table 10: Reproduction accuracy for polyglot TTSs

	1 Word Error	2 words Errors	More than 2
English	67	17	10
Spanish	98	2	0
Malay (en1)	45	35	20
Malay (af1)	18	23	59

In the reproduction test, the participants were asked to retype what they thought they heard from the short sentences using the polyglot synthesised speech.

Table 11: Respondents' opinion on the quality of the polyglot TTSs

	Very Good	Good	Fair	Poor	Very Poor
English	4	22	39	27	8
Spanish	0	4	31	61	4
Malay (en1)	5	6	28	38	23
Malay (af1)	0	0	12	23	65

The respondents feedback on the effort required to understand the synthesised speech is reflected by the quality rating.

Finally, the respondents were asked to compare the polyglot TTS to the monolingual TTS. The results show that the English and Malay(both TTSs) monolingual TTSs are not significantly preferred than the polyglot synthesis. These results are unexpected since it was assumed the monolingual TTSs would be strongly preferred for all languages.

Table 12: Respondents effort to understand the polyglot TTSs

	Least Effort	Mod. Effort	Consider. Effort	None Understood
English	19	42	31	8
Spanish	37	41	20	2
Malay (en1)	18	30	41	11
Malay (af1)	1	12	46	41

Table 13: Respondents preference of the polyglot TTSs over the monolinguals

	Better	Slight. Better	Abt. the Same	Slight. Worse	Worse
English	13	14	21	31	21
Spanish	4	0	16	17	63
Malay (en1)	7	23	33	30	7
Malay (af1)	10	21	27	23	19

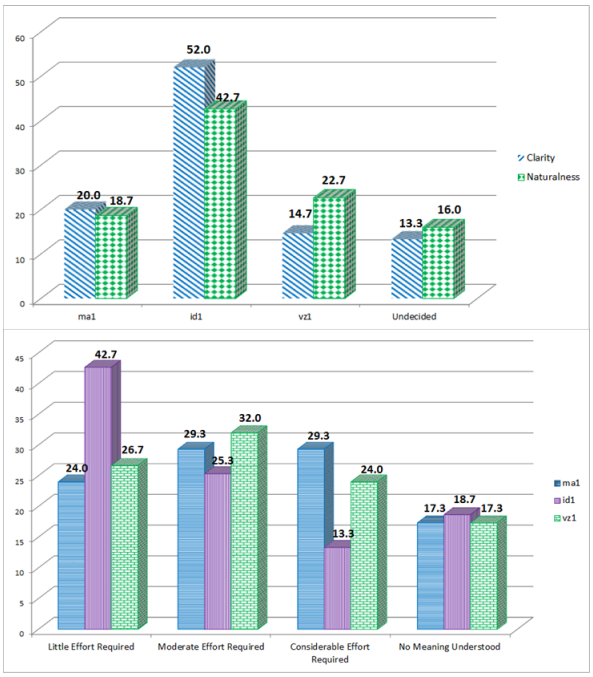

Figure 2: General Quality and Naturalness Rating for Initial Iban TTS Evaluation

5.2. Iban Results

The Iban TTSs were evaluated in two parts - the first evaluated the understanding of complete meaningful sentences while the second used nonsense sentences. Similar to the previous surveys, respondents were first required to give their feedback on the synthesised speech of five complete sentences. Each sentence was generated using three polyglot synthesisers from three different resources. These are from Malay(ma1), Indonesian(id1) and Spanish(vz1). The TTS using Malay resources represented a benchmark TTS and was built by mimicking natural speech. It was compared to the Iban polyglot TTS using Indonesian resource that adapted the prosody from open intonation language and the polyglot TTS using Spanish resources that adapted the prosody from the stressed language. During the second part of the evaluation, the respondents were requested to type back nonsense sentences synthesised by two polyglot TTSs from Malay and Indonesian resource. Fifteen Iban native speakers participated in this evaluation.

Figure 2 shows the results of the first part of the evaluation. Respondents were asked to rate the general quality, naturalness and effort required to understand the synthesised speech. The values given are in percentage. After the respondent completed the first part of the survey, they were requested to do a perception test. Figure 3 shows the detail perception results. The respondents were asked to type back what they thought they heard (results in Figure 3: **Accuracy**)and then rate the quality(results shown in Figure 3: **Quality**). The respondents were also asked about the effort they perceived they had to make to understand the speech (Figure 3: **Effort**). Finally the respondents were asked to rate which one of the two synthesised speech the respondents more preferred (Figure 3: **Preference**). As Figure 3 shows, the Indonesian-based TTS is highly preferred to the Malay-based TTS.

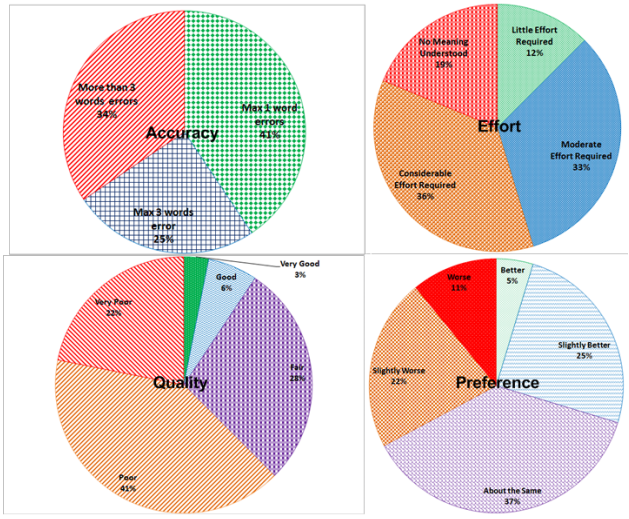

Figure 3: Results for Nonsense Iban Sentences

6. Discussion and Conclusion

This paper presented an approach to adapting available resources to create TTSs for resource poor languages. To evaluate our approach, TTSs for two resource rich languages were implemented, namely English and Spanish. These polyglot TTSs are compared to a monolingual TTS. While, as expected, the overall quality of the polyglot TTS were rated lower than their corresponding monolingual TTS, all polyglot TTS produced acceptable quality speech synthesis and in some cases were preferred by users to their corresponding monolingual TTS. TTSs for two resource poor languages, Malay and Iban, were also constructed. The Malay polyglot TTS was developed using English and Afrikaans data and the Iban polyglot TTS was

developed using Malay, Indonesian and Spanish data.

We have demonstrated that our approach is capable of producing acceptable polyglot speech synthesis from different language data. Although the output is not as good as a monolingual TTS for English, Spanish and Malay, the polyglot synthesised speech is acceptable to native speakers. Considering the polyglot TTSs for English, Spanish and Malay all follow the same generic prosody representation, and are generically designed, we believe that with language specific refinement, the TTS can be further improved.

To test our approach on a resource poor language, we developed three TTS for Iban using Malay, Indonesian and Spanish as resources. The Iban TTS from Malay voices is constructed by mimicking native speech. The original Indonesian and Spanish voices for the Iban TTSs are constructed based on the proposed prosodic representation, but one used a non-stressed and non-tonal prosody while the other used a stressed language prosody.

Although the Malay resources was used to mimic the Iban native speakers' speech, the respondents rated the TTS built using Indonesian resources as more natural and more clear than the Malay-resource-based TTS. Also, when the rating for Spanish resource is compared against the Malay resource, the respondents find that the naturalness is better when using Spanish resources but worse than when using Malay in term of quality. Future research will further investigate the factors behind the relative effectiveness of different related language resources.

7. Acknowledgements

The authors acknowledge the contributions of the researchers from Sarawak Language Technology (SaLT) Research Group at Universiti Malaysia Sarawak (UniMAS). Special appreciation goes to Sarah Flora Samson Juan for preparing the Iban phoneme set. Authors also would like to thank Ismail Bhula and Phil Smith from the School of Computer Science, University of Birmingham for helpful feedback on early drafts of this paper.

8. References

David Crystal. 2010. *The Cambridge Encyclopaedia of Language*. Cambridge University Press, Cambridge, 3rd ed. edition.

Zuraidah Don, Gerry Knowles, and Janet Yong. 2008. How Words can be Misleading: A Study of Syllable Timing and "Stress" in Malay. *The Linguistics Journal*, 3(2).

Daniel Hirst. 2005. Form and Function in The Representation of Speech Prosody. *Speech Communication*, 46(3-4):334–347.

Daniel Hirst. 2007. A Praat Plugin for MOMEL and INTSINT with Improved Algorithms for Modelling and Coding Intonation. In *Proceedings of ICPhS 2007*, pages 1233–1236, Saarbrücken, Germany, 6-10 August 2007. Universität des Saarlandes.

M. Paul Lewis. 2011. Iban: A Language of Malaysia (Sarawak). Internet. Accessed on December 2011.

Jonathan Marks. 2011. *English Pronunciation in Use*. Cambridge University Press, Cambridge, 3rd ed. edition.

Christian Maynes and Farrel Gara. 2011. Useful Iban Phrases. Internet. Accessed on December 2011.

MBROLA-Group. 2005. The MBROLA PROJECTS Towards a Freely Available Multilingual Speech Synthesizer. Internet. Accesed on 17th July 2009.

PhoTransEdit-Group. 2011. Text to Phonetics. Internet. Accesed on 10th November 2010.

Suhaila Saee. 2011. Apa itu Jako Iban? Email correspondence, September.

Keiichi Tokuda, Heigen Zen, and Alan W. Black. 2002. An HMM-Based Speech Synthesis Applied to English. In *Proceeding of IEEE Workshop on Speech Synthesis 2002*, pages 227–230, September.

Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner. 1999. From Multilingual to Polyglot Speech Synthesis. In *Proceedings of Eurospeech 1999*, pages 835–838, Budapest, September.